3O1-2in

実世界における身体動作のコーディング・セグメンテーション手法の提案

Suggestion the method for coding and segmentation of body movements at the filed

牧野 遼作[*1] 城 綾実[*2] 坊農 真弓[*1,2] 高梨 克也[*3] 佐藤 真一[*2] 宮尾 祐介[*1,2]
Ryosaku Makino Ayami Joh Mayumi Bono Katsuya Takanashi Shin'ichi Satoh Yusuke Miyao

[*1] 総合研究大学院大学複合科学研究科 [*2] 国立情報学研究所コンテンツ科学研究系 [*3] 京都大学学術情報メディアセンター
School of Multidisciplinary Science, The Graduate University For Advanced Studies (SOKENDAI) #1 National Institute of Informatics #2 Academic Center for Computing and Media Studies, Kyoto University#2

This study aims to determine the method for coding and segmentation of body movements in the field. In terms of communication encoding or segment body movement, two issues occurred while conducting the research: (1) The content of coding body movement did not coincide between workers and (2) The onset and end time of body movement did not coincide. The latter issue was resolved by use of the method described in this study. This method involved splitting video recordings of body movement into one-second data segments. The one-second increments are referred to as windows. The researcher divided the windows into the following categories: Whether movement was detected, whether movements were intended, and whether the lack of movement was significant for communication. This method allowed for use of the appropriate time framework for classification of body movement consistent with the purpose of the research.

1. はじめに

筆者らは，実世界における人々のコミュニケーションを異分野融合型のアプローチで解明していくための環境構築を試みている．具体的には，これまで日本科学未来館の科学コミュニケーターの実践についてフィールド調査を行い[坊農 13, Bono 14]，そこで得られたデータをマルチモーダルコーパスとして整備する試みを進めてきた[城 14，城, 15]．

本稿では，構築中のマルチモーダルコーパスの中で，特に身体動作のコーディング・セグメンテーションの一手法を紹介する．本稿では，まず，マルチモーダルコーパス作成のために，なぜ身体動作のコーディング・セグメンテーション手法の開発が必要なのかを述べ，続けて本手法を検討するために用いたデータについて述べる．次に，提案を行うコーディング・セグメンテーション手法について詳述し，本手法の実施事例の分析結果を述べる．最後に，本手法による身体動作のコーディング・セグメンテーションを利用した異分野融合研究の展望について述べる．

1.1 背景・目的

人と人とのコミュニケーションの研究は様々な研究領域(e.g. 言語学，心理学，認知科学，社会学，工学)で行われている．近年では，相互行為中で産出される音声発話だけではなく，身体動作への関心が高まっている[Streeck et.al 11]，Wachsmuth 08]．これまで，コミュニケーションにおける身体動作は，どのような形態を持つのか，どのような機能を持つのか，何を指し示したものなのか，などの観点によって分類されたり[McNeill 92，Kendon 04]，単に動作の自体を記述する(e.g. 前方へ右手の人差し指を伸ばす)だけではなく，人々にとって報告可能な形(e.g. 展示物を指差す)で記述されたりしてきた[西阪 07]．しかしながら，これまで，動作の分類・記述方法に，一般的な方法は存在しなかった．これは，分類・記述ともに，それぞれの研究目的に適切な形でなされてきたからであると指摘されている[Mondada 07]．つまり，動画データの中では，多様な身体動作が観察可能であるが，その中から，どのような身体動作を研究の対象とするかは，研究目的に従うものであり，また動作の分類・記述の方法の決定も，研究目的によるものであるということである．しかしながら，長時間におよぶデータの全てから，研究対象とする身体動作の適切な事例を効率的に収集することは困難である．そこで，本稿では，動画データの中から，様々な研究目的に沿ってなされる身体動作を記述・分類するための，基準となる身体動作の開始点・終了点を提供する．

連絡先：牧野遼作，総合研究大学院大学，ryosaku@nii.ac.jp

1.2 日本科学未来館 SC データ

本稿では，「日本科学未来館 SC 動画データ(以下，未来館データ)」ついて説明する．この未来館データは，筆者らが整備を行っている「日本科学未来館 SC コーパス」[城 15]に含まれる動画(mov 形式)データである．このコーパス・データでは，日本科学未来館に所属する科学コミュニケーター(以下 SC)が来館者に展示物の説明を行う場面を収録したもので，35 事例，合計約 8 時間からなる．このデータは，SC 自身が日常業務の中で行っている複数の展示空間での展示物の解説場面を収録したもので，自然な環境での多様な身体動作が含まれている．そのため，本稿の提案手法は，未来館データ以外の他のデータ・セットに対しても適用可能なものとなると考えられる．

2. セグメンテーション・コーディング手法

本稿で提案する手法は，2段階から構成される．第 1 段階は，動画を1秒刻みの幅で分割するものである．第 2 段階は，1秒刻みの幅の中に，設定された注釈を選択，記入していく．以上の作業に基づき，研究目的に沿った注釈を選択し，動作の開始点・終了点を抽出が可能である．今後の利用としては，後述のように，この開始点・終了点を基準に，動作の分類・記述を行っていくことが可能である．図 1 は，本手法の大まかの概要を示したものである．

2.1 第 1 段階: ウィンドウの設定

基準となる時間的な枠組みを提供するためには，動作の開始時間と終了時間を同定する必要がある．1 フレームごとに再生可能な動画が収録可能となった現代において，様々な動作

の開始点・終了点は，1 フレームごとに厳密に決定することが可能となった．このように身体動作が記述可能になったことにより，一つの発話と，それに伴った現れる身体動作の関係性が検討されてきた[McNeill 2005]．一方で，1 フレームごとに決定しようとする開始点・終了点は，熟達した作業者間でも，ズレが生じてしまうものである．このズレを解消しつつ，長時間のデータから，身体動作の適切な事例を効率的に収集するためには，身体動作の 1 フレームごとの厳密な開始点・終了点を決定することは，非効率である．そこで，本手法では動画全体を 1.0 秒刻みの幅で分割を行った(以下，ウィンドウ)．このウィンドウ内に，注釈内容を選択していくことによって，複数の作業者間での動作の開始点・終了点の細やかなズレが生じないことになる． このウィンドウに対して，下記の注釈層ごとに，注釈を選択していくことで，動画データのセグメンテーション・コーディングを安定的かつ効率的に実施できる．

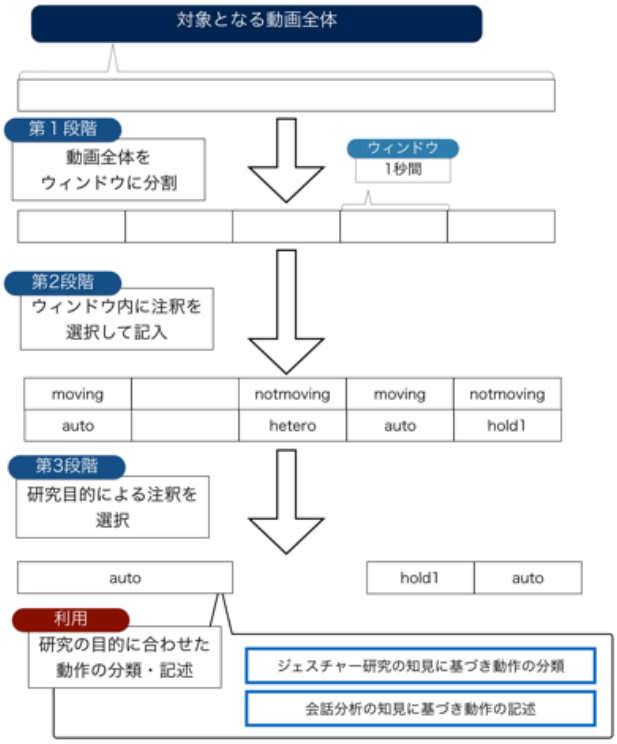

図 1 本手法の概要

て困難なときは unclear を，対象参与者の手が動画内で確認できないときには missing を選択する．

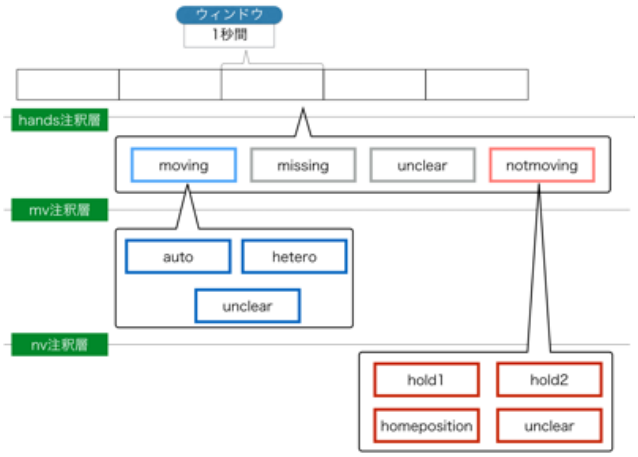

図 2 本手法の注釈と注釈層

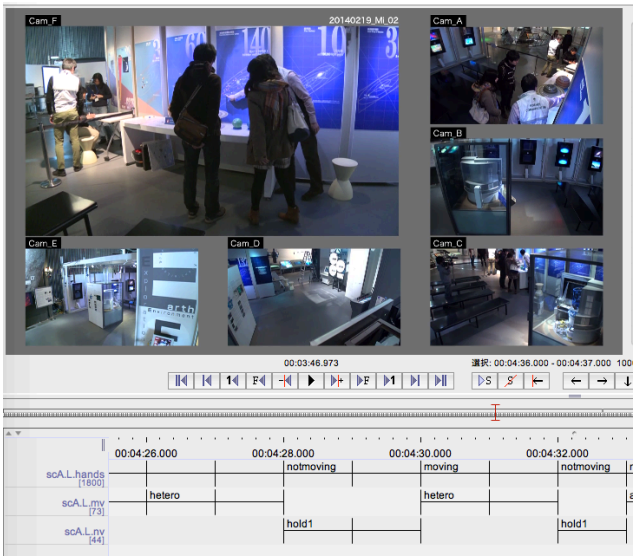

図 3 ELAN 上での本手法の作業状態

2.2 第 2 段階: 注釈層への注釈の選択・記入

本研究で提案する手法は，ELAN[1]上で作業を進めながら検証を進める．ELAN は動画を再生しながら，動画内容の記述を行うことができるソフトで，複数の注釈層を設定することができる．本手法では，参与者ごとに hands，mv，nv の 3 層×右/左手の計 6 層を設ける (図 2)．また各注釈層では，それぞれいくつかの注釈内容を選択することができる．以下，注釈層と注釈内容の説明を記す(図 3)．

(1) hands 注釈層

hands 注釈層では，対象参与者の手が動いているか否かを注釈するものである．参与者の手がウィンドウ内で明らかに動いていた場合は moving を選択し，逆に明らかに動いていないときには notmoving を選択する．また，この判断が作業者にとって困難なときは unclear を，対象参与者の手が動画内で確認できないときには missing を選択する．

(2) mv 注釈層

mv 注釈層では，hands 注釈層で moving と選択された対象参与者の手の動きを分類する．ジェスチャーや実用的な動きなど，対象参与者が自発的に動かそうとした手の動きの場合は auto，歩行や前の手の動きの反動などによる動きは hetero を，この判断が作業者にとって困難なときは unclear を選択する．

(3) nv 注釈層

nv 注釈層では，hands 注釈層で，notmoving と選択された対象の参与者の停止状態を home position，hold1，hold2，unclear の4種類に分類する．

停止状態は，ジェスチャー研究の知見に基づき，ホームポジション(home position)[Sack 2001]とホールド(hold)[Kita 94]の 2 種類に分けられる．ホームポジションとは，ジェスチャー動作開始前と終了後に置かれる手の位置のことを指し，リラックスポジション(relax position)とも呼ばれる[Kendon 04]．本手法では，対

[1] https://tla.mpi.nl/tools/tla-tools/elan/

象者の手が自身の腰より低い位置にある状態で，動作を終えてリラックスした状態であるときに，nv 注釈層で home position を選択する．停止状態のもう一つのホールドとは，一連の動作の最中の停止状態である．例えば，人が指差しを行うポインティングジェスチャーは，動作が開始され，何かを指差された状態で停止する．このとき，指差し状態で止まった手は一連のジェスチャーを構成する"動作"の一部であり，ホールドと呼ばれる．本手法では，対象者の手が自身の腰より上の高さにある状態で，前の動作の途中で一時的に停止したものと判断された場合は，hold1 を選択する．

一方で会話の中で手の停止状態に関する研究では，既存のホールドのように，ジェスチャーの構成要素となっているわけではないが，ホールドとは異なるコミュニケーションに対する機能を持っている現象が報告されている．例えば，細馬[08]は，一連のジェスチャー終了後に，次のジェスチャーを開始するまでの間に生じる hold によって，自身の発話がさらに続くことが示されるという事例を報告している．また腕組みを行っている人は，自身の内面感情を表出したり[大坊 98]，自身が相手の話を聞いていることを示している可能性がある． このことから，腰より高い位置の停止状態であり，単純には hold1 が選択される停止状態であっても，ホールドではない可能性があり，また今後のコミュニケーション研究において，研究対象となる可能性がありうる． そこで，本手法では，手が参与者自身の腰より上の高さにある状態で，前の動作が終わった後の停止の場合，一時的に停止したものであると判断された場合は，hold2 を選択し，既存のジェスチャー研究におけるホールドとは区別し，hold1 と hold2 の分類を行った．これらの注釈の記入は，以下のように行う．hands 注釈層の事前に設定されたウィンドウに注釈を選択していく．このとき，注釈状態が変更されたウィンドウにのみ注釈を記入していく (e.g. 動作の開始が含まれたウィンドウには moving を選択する．動作が続く限り，以降のウィンドウは空欄とする．次の注釈状態変化したウィンドウ，すなわちウィンドウ内で身体が停止状態なったときのウィンドウに notmoving を選択し記入する)． mv 注釈層では，hands 注釈層で moving とされた区間(moving が選択されたウィンドウから，他の注釈が選択されたウィンドウの一個前まで)の中で，mv 注釈層に注釈を選択し記入し，notmoving となされた区間に nv 注釈層に注釈を選択して記入する．このとき，hands 注釈層と同様に注釈状態の変更されたウィンドウにのみ注釈を記入すればよい．

2.3 第 3 段階: 研究目的による注釈の選択・

第 3 段階として，以上のようにウィンドウに記入された注釈の中から，研究目的に則した注釈を取り出すことによって，動作の開始点・終了点，時間的な枠組みを抽出することができる．例えば，コミュニケーションの中でジェスチャー(さらにそれに類する動作)を取り出そうとするのであれば，mv 注釈層の auto と nv 注釈層の hold1 を取り出すことによって，動画の中から分析対象を絞り込むことが可能である．図 4 は，本稿の手法の結果より，選択した注釈(auto と hold1)の開始点から終了点までを ELAN 上でつなげたものである．この auto と hold1 と記入された注釈の中に，動作の分類や記述を記入することができる．

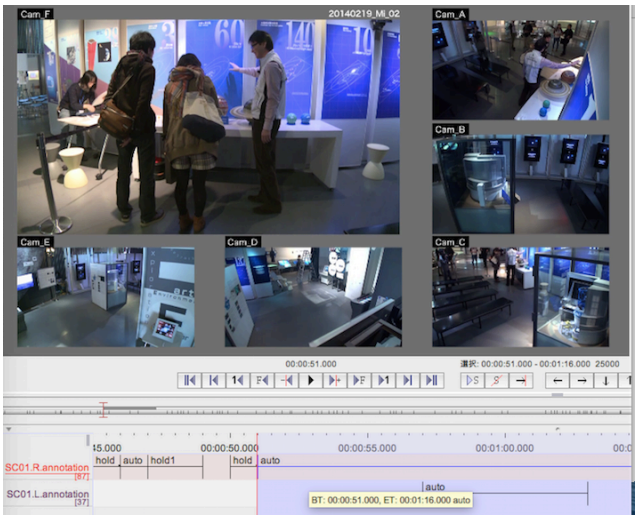

図 4 ELAN 上での身体動作の開始点・終了点

3. 分析

本稿で提案する手法の有用性を検定するために，まず，コーディング・セグメンテーションの信頼性について，分析を行う．

3.1 方法

(1) 対象データ

SC は来館者と比べて，身体動作の産出頻度が多く観察されたため，コーディング・セグメンテーションは SC の身体動作に対してなされた．さらに，SC の個人差を考慮するため，異なる SC が参与しているデータを2つ選び，その中からランダムに抽出した2分間をコーディング・セグメンテーション対象とした．

(2) 作業者

コーディング・セグメンテーションの作業者は，手法について説明を受けた3名である．

(3) 分析手法

上記のコーディング・セグメンテーションの作業結果より， mv 注釈層の auto の注釈と，nv 注釈層の hold1 の注釈を対象に，3名の作業者間の一致率について検定を行った． mv 注釈層の auto と nv 注釈層の hol1 は，それぞれジェスチャーとそれに類する身体動作／身体の停止状態である．これらは，今後相互行為分析による記述のベースになるものであり，まず一致率を検討するべきものと考えられる．

3.2 結果と考察

一致率検定の結果，K=0.748 となり，3 名の作業者間のコーディング・セグメンテーションは，実質的に一致していると言えた．また，作業対象となる SC の右手・左手の動きを別々に信頼度検定を行った結果，右手は K=0.805，左手は K=0.675 となった．このことから，本稿の方法は，右手(利手)のように，動作が産出しやすいものに対しては特に高い一致率を示す可能性が示唆された．

4. おわりに

最後に，本稿の手法を用いて提示する時間的な枠組みの複数の領域での利用について，将来展望を述べる．

4.1 会話分析・相互行為分析

会話分析・相互行為分析では，コミュニケーションの中で起こる現象の記述を行う．このとき，発話の場合，記述対象の開始点・終了点は，順番[Schegloff 74]をベースに決定される．しかしながら，身体動作については，開始点・終了点を決定することは難しい．本稿の手法は，記述対象となる身体動作の開始点・終了点を決定するための一つの材料となりえるだろう．ただし，本稿で提案される身体動作の開始点・終了点間の幅はかなり広いものである．そのため，記述レベルの粒度によっては，複数の身体動作が統合された開始点・終了点となっていると見なされるかもしれない．

特に，mv 注釈層の auto 注釈については，さらなる検討が必要である．auto 注釈では，対象者が自発的に動かそうとした手の動かせている状態である．だが，auto 注釈が連続していたとしても，それが一つの連続した動作とは限らない．例えば，参与者があるジェスチャーA を行った後，ホームポジションを介さず，他のジェスチャーB を連続して行ったとき，本手法で提案される動作の開始点・終了点の枠組みの中に，2つのジェスチャーが含まれているといえる．これらの2つのジェスチャーは研究目的によって，動作の記述をどのような形で行うかによって，どのように切り分けられるかが決定される．単に，動作を「ジェスチャーしている」と記述するだけならば，本手法が提供する動作の開始点・終了点の枠組みで構わないだろう．しかし，ジェスチャーA が指差しジェスチャーであり，ジェスチャーB が展示物の説明のためのジェスチャーをしているときならば，この2つのジェスチャーを一つの枠組みで「ジェスチャーしている」と記述しているのは，不十分であるかもしれない．これらのジェスチャーを切り分けるには，ジェスチャー研究の知見において，ジェスチャーは“ジェスチャー単位”と呼ばれる単位[Kendon 04, McNeill 05]などをどのように反映するかを検討する必要がある．

4.2 自然言語処理・画像処理

本稿で提案される手法は身体動作の開始点・終了点を決定するものであり，開始点・終了点の中に，どのような分類・記述を入れていくかについては，研究目的によって様々なものを入れることができる．例えば，主にジェスチャー研究における身体動作の分類コーディングを記入することも可能である．未来館データでは，Kinect による収録がなされているため，身体動作のカテゴリーと Kinect により得られた三次元映像・骨格情報とを統合した画像処理研究も可能であろう．

また，時間枠組みの中に，会話・相互行為分析の記述を複数の研究者が行っていき，そのデータを学習用データとすることによって，身体動作を表す記述を自動生成する自然言語処理分野での研究も可能となると考えられる．

参考文献

[坊農 13] 坊農真弓・高梨克也・緒方広明・大崎章弘・落合裕美・森田由子: 知識共創インタフェースとしての科学コミュニケーター: 日本科学未来館におけるインタラクション分析, ヒューマンインタフェース学会論文誌, Vol.15, No.4, pp.375-388, 2013.

[Bono 14] Bono,M., Ogata,H., Takanashi,K., and Joh,A. : The Practice of Showing ‘Who I am’: A Multimodal Analysis of Encounters between Science Communicator and Visitors at Science Museum, Universal Access in Human-Computer Interaction, Universal Access to Information and Knowledge Lecture Notes in Computer Science, Vol8514, pp. 650-661, 2014.

[大坊 98] 大坊郁夫: しぐさのコミュニケーション―人は親しみをどう伝えあうか, サイエンス社,1998.

[細馬 08] 細馬 宏通: 話者交替を越えるジェスチャーの時間構-隣接ペアの場合-,認知科学,vol16(1),pp.91-102, 2008.

[城 14]城綾実・牧野遼作・坊農真弓・高梨克也・佐藤真一・宮尾祐介: 異分野融合によるマルチモーダルコーパス作成-展示フロアにおける科学コミュニケーションに着目して-. 人工知能学会言語・音声理解と対話処理研究会資料, SIG-SLUD-B401, 2014.

[城 15] 城綾実・牧野遼作・坊農真弓・高梨克也・佐藤真一・宮尾祐介: 異分野融合によるマルチモーダルコーパス設計-各種アノテーション方法と利用可能性について-, 言語処理学会第 21 回年次大会, pp.561-564, 2015.

[Kendon 04] Kendon,A. : GESTURE Visible Action as Utterance. CAMBRIDGE UNIVERSITY PRESS, 2004.

[Kita 90] Kita,S. : The temporal relationship between gesture and speech, A study of Japanese-English bilinguals, 1990.

[McNeill 92] McNeill,D. : Hand and mind, The University of Chicago Press, 1992.

[McNeill 05] McNeill,D. : Gesture and Thought, The University of Chicago Press, 2005

[Mondada 07] Mondada, L. : Commentary: Transcript variations and the indexicality of transcribing practices, Discourse Studies, vol9(6), pp.809-821, 2007.

[西阪 08] 西阪仰: 分散する身体―エスノメソドロジー的相互行為分析の展開, 勁草書房,2008.

[Sacks 02] Sacks, H. and Schegloff, E.A: Home position, Gesture, vol2(2), pp.133-146,2002.

[Schegloff 07] Schegloff, E. A. : Sequence Organization in Interaction: A Primer in Conversation Analysis, Cambridge: Cambridge University Press, 2007.

[Streeck 11] Streeck,J., Goodwin,C., and LeBaron,C. : Embodied Interaction: Language and Body in the Material World, Cambridge University Press, 2011.

[Wachsmuth 08] Wachsmuth,L., Lenzen,M., and Knoblich, G. : Embodied Communicaiton in Human and Machines, Oxford University Press,2008.